www.bsc.es

Paris, 19 April 2016

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Data for climate and air quality

Francisco J. Doblas-Reyes
BSC Earth Sciences Department

iCrea

RDA Research Data Sharing without barriers
RESEARCH DATA ALLIANCE

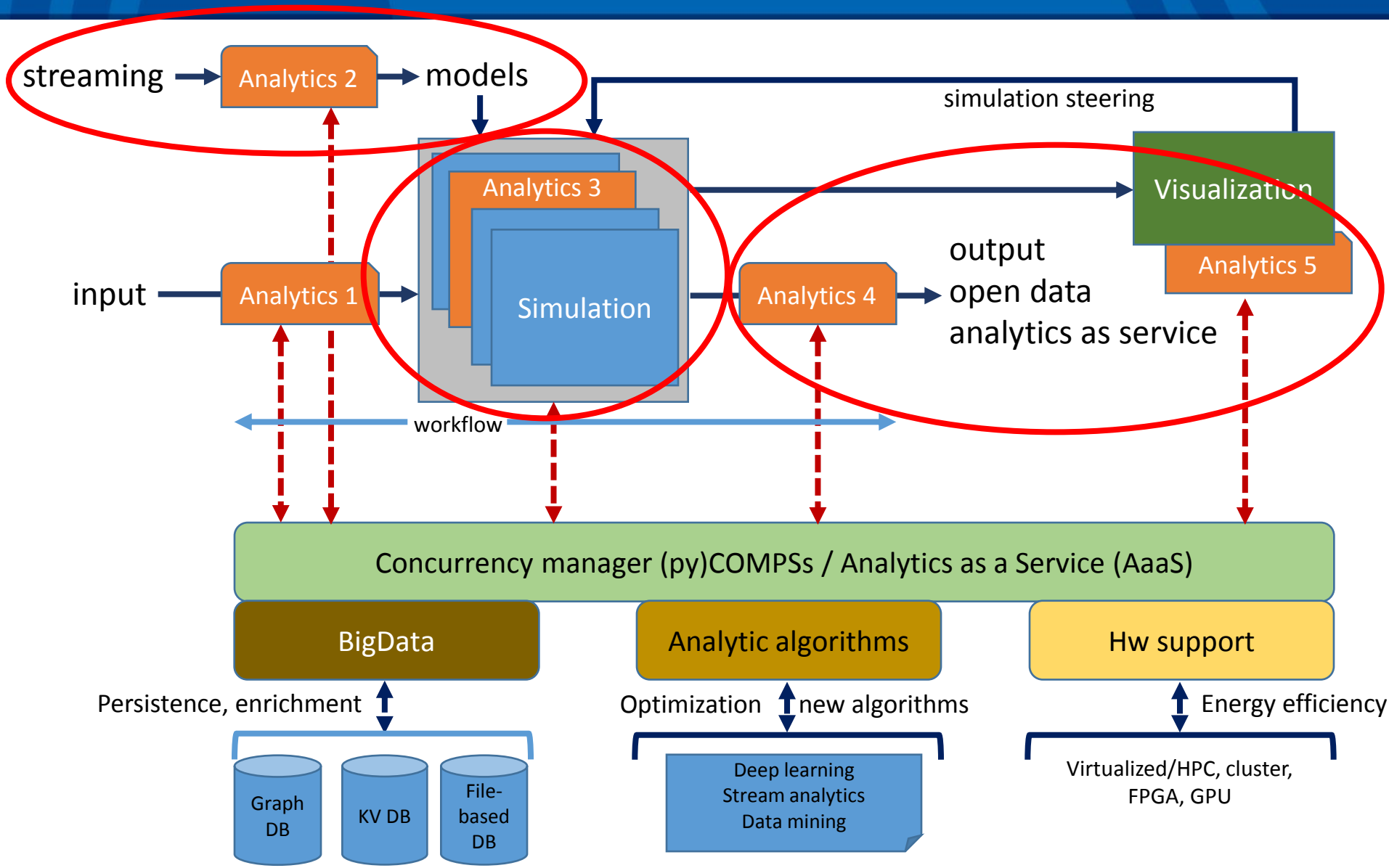
YouTube

Data in our field

- There are problems involving large, complex datasets: operational weather and air quality prediction.
- There are large problems involving data: simulation of anthropogenic climate change.
- And there are Big Data problems: dealing with heterogeneous data sources to produce end-user information with a weather, climate and air quality component.

I will not address the issue of open data and assume that access to data is not a difficulty (a strong assumption).

A conceptual model from CS



Case 1: Data streaming for air quality forecasting

Case 2: Simultaneous analytics and HPC in climate prediction

Case 3: Analytics as a service

CALIOPE air quality operational forecasts

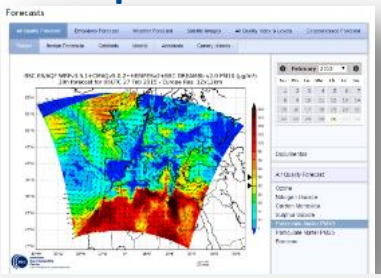
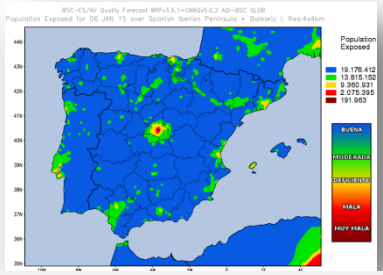
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

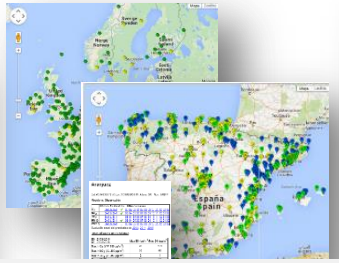
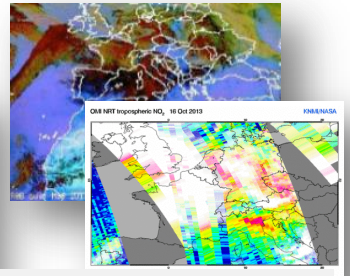
EXCELENCIA SEVERO OCHOA

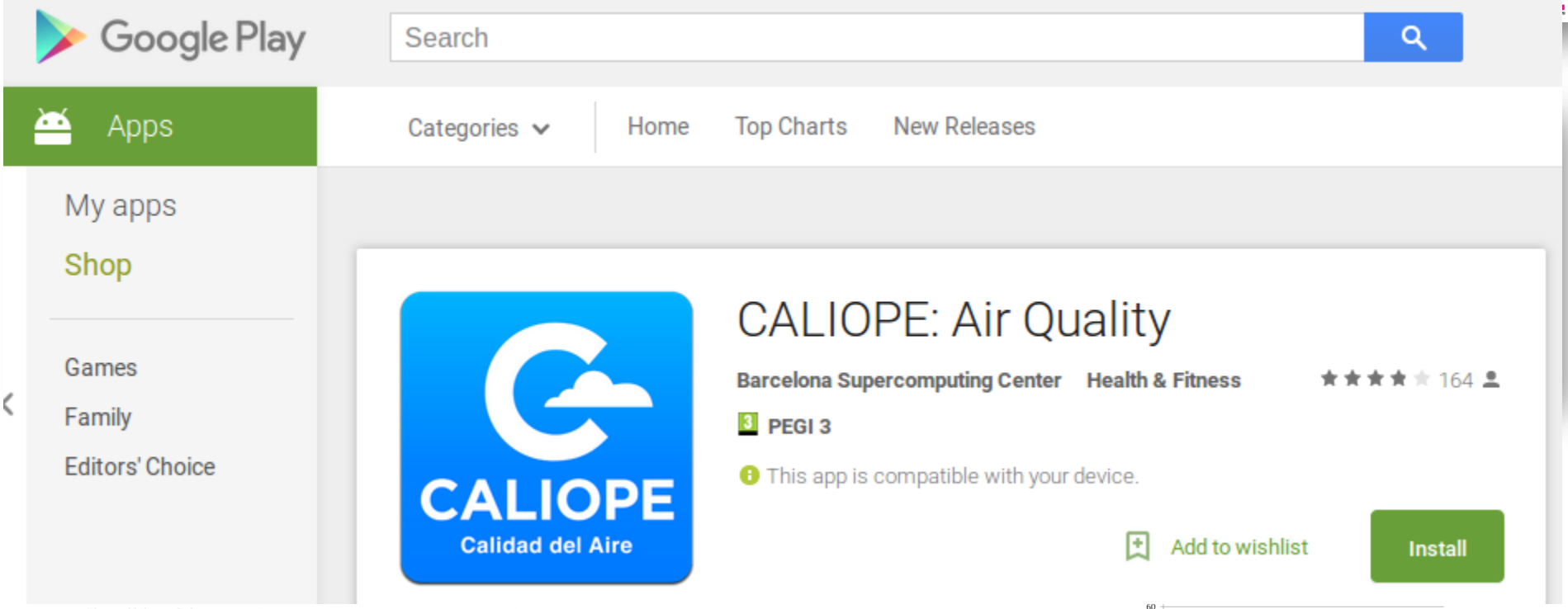
AQF CALIOPE system: daily forecast and evaluation

Daily forecast for **meteorology, emissions and air quality**: Europe (12km), Iberian Peninsula (4km), Andalusia, Catalonia and Madrid (1km), since 2007

Forecast products



PM10_KF annual average skill evolution (2011-2014)

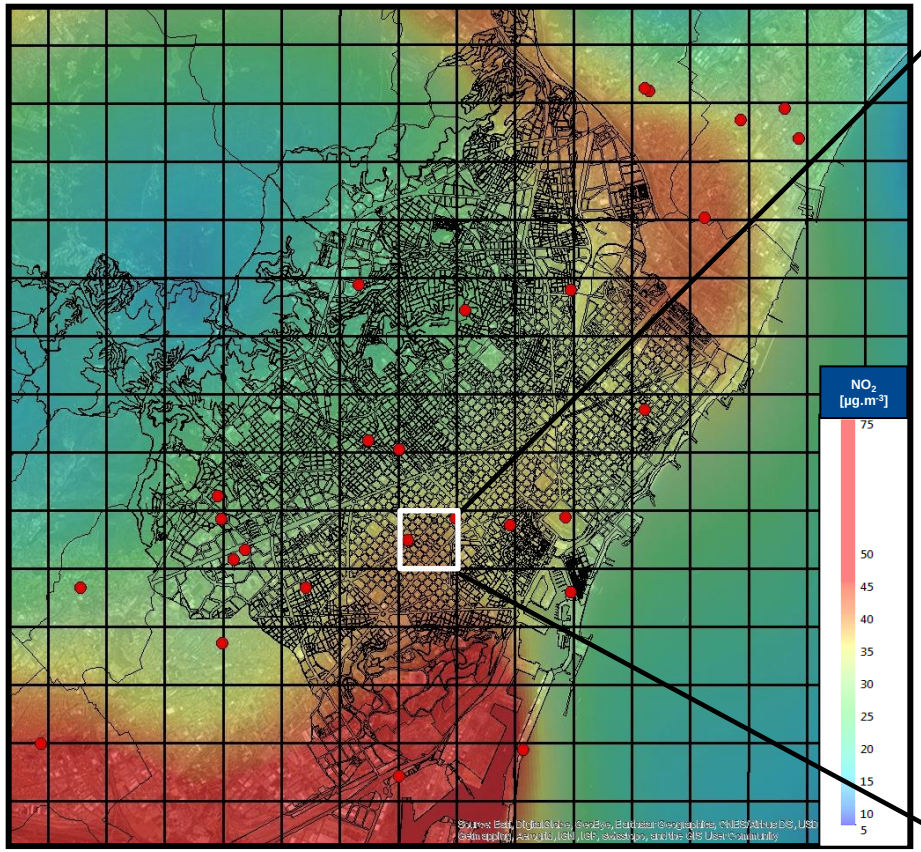
New air quality management tools

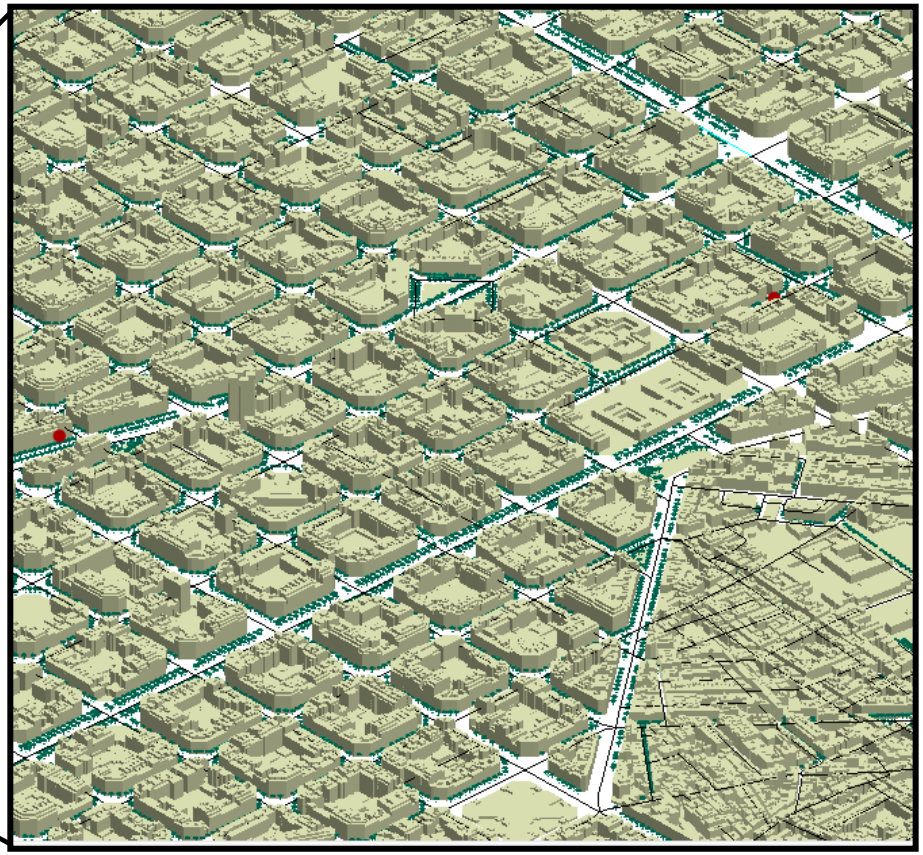
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Where we are now

Where we want to be





Objective	To develop an air quality model based on a CFD coupled to an atmospheric chemistry model at city scale, enhanced by the use of Big Data technologies, to assess urban air quality.

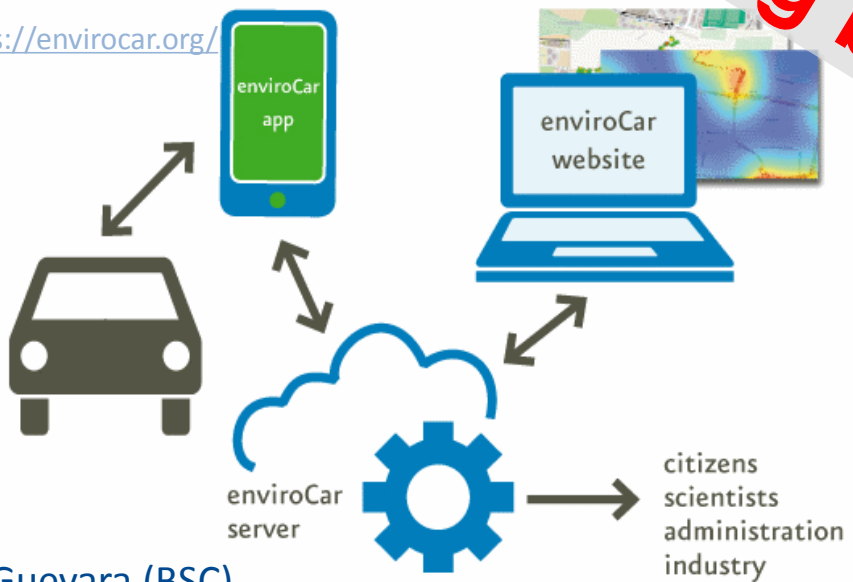
Data streaming for emissions

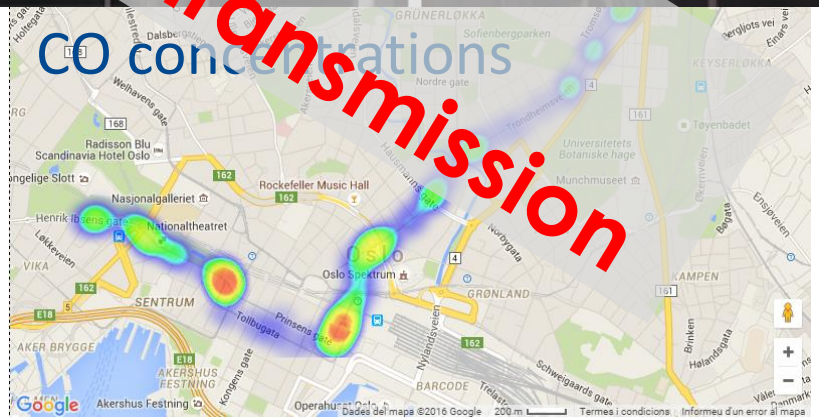
High-resolution air quality modelling requires appropriate emissions

- Collection and processing of sensor-generated data to feed a real-time emission model (and to validate air quality predictions)
- Providing the sensors with the adequate technology is a challenge
- Managing large volumes is another one: sampling 10 Hz, ten variables ~30 MB/day, city-wide ~300 GB/day (10,000 vehicles)

Many more sensors possible
Need for computing before transmission

https://envirocar.org/





M. Guevara (BSC)

Case 1: Data streaming for air quality forecasting

Case 2: Simultaneous analytics and HPC in climate prediction

Case 3: Analytics as a service

Running global climate predictions

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
EXCELENCIA SEVERO OCHOA

Climate prediction allows running jobs independently and simultaneously by wrapping together ensemble members for different start dates. This is not trivial parallelisation.

A workflow manager is required.

5,500 simulated years -> 1.4 PB

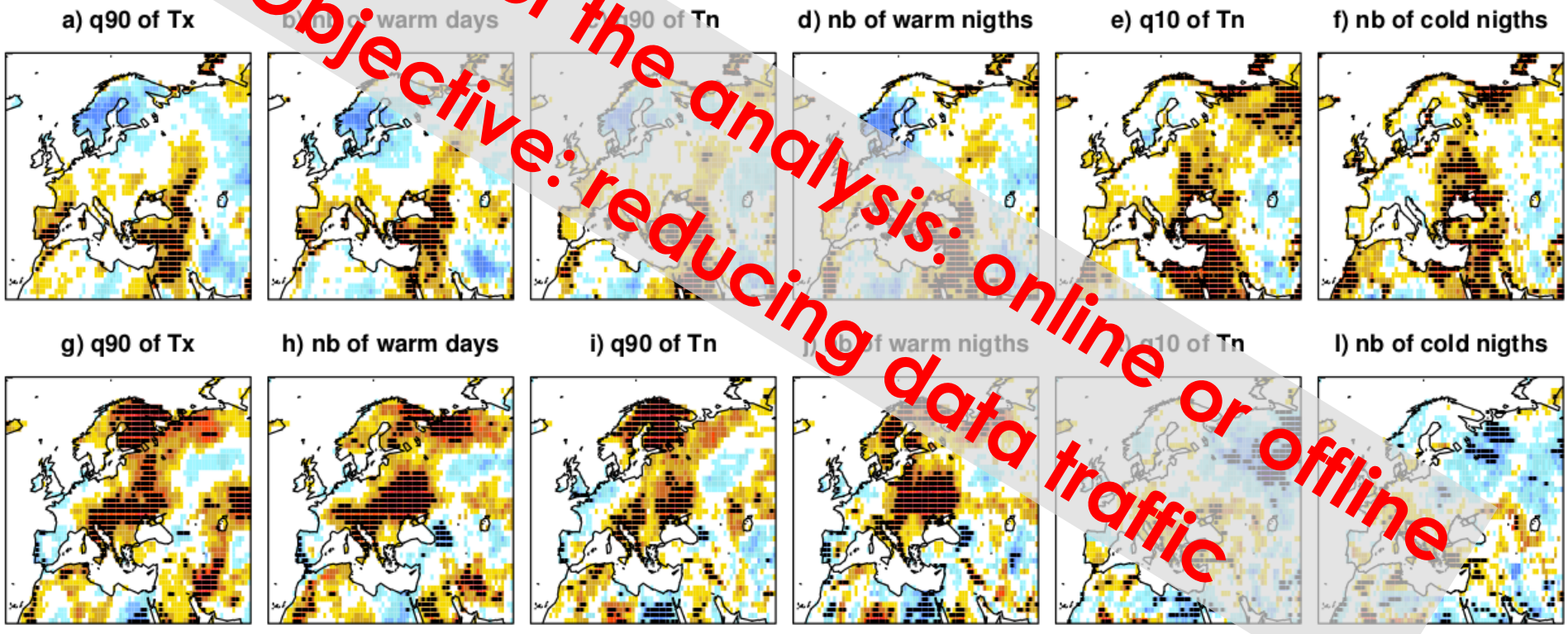
EC-Earth 3 at Lindgren, PDC						
Number of Start Dates		1	5	10	10	20
Number of Members		1	5	5	10	10
Number of Independent Simulations		1	25	50	100	200
T159-ORCA1	Cores	1440	3600	7200	14400	28800
	Wall-clock Time (Hours) / year	5	5	5	5	5
	CPU Time (Hours) / year	720	18000	36000	72000	144000
	Output Size (GB) / year	10,80	480	960	1920	3840
T255-ORCA1	Cores	360	9000	18000	36000	72000
	Wall-clock Time (Hours) / year	5	5	5	5	5
	CPU Time (Hours) / year	1800	45000	90000	180000	360000
	Output Size (GB) / year	19,20	5184	10368	20736	41472
T799-ORCA025	Cores	1104	27600	55200	110400	220800
	Wall-clock Time (Hours) / year	40	40	40	40	40
	CPU Time (Hours) / year	44160	1104000	2208000	4416000	8832000
	Output Size (GB) / year	256,80	6420	12840	25680	51360

Predicting extremes

JJA near-surface temperature correlation of the ensemble mean from experiments with a climatological (top) and difference with one with realistic (bottom) land-surface initialisation. Results for EC-Earth2.3 started in May over 1979-2010.

Two ways for the analysis: online or offline

Objective: reducing data traffic

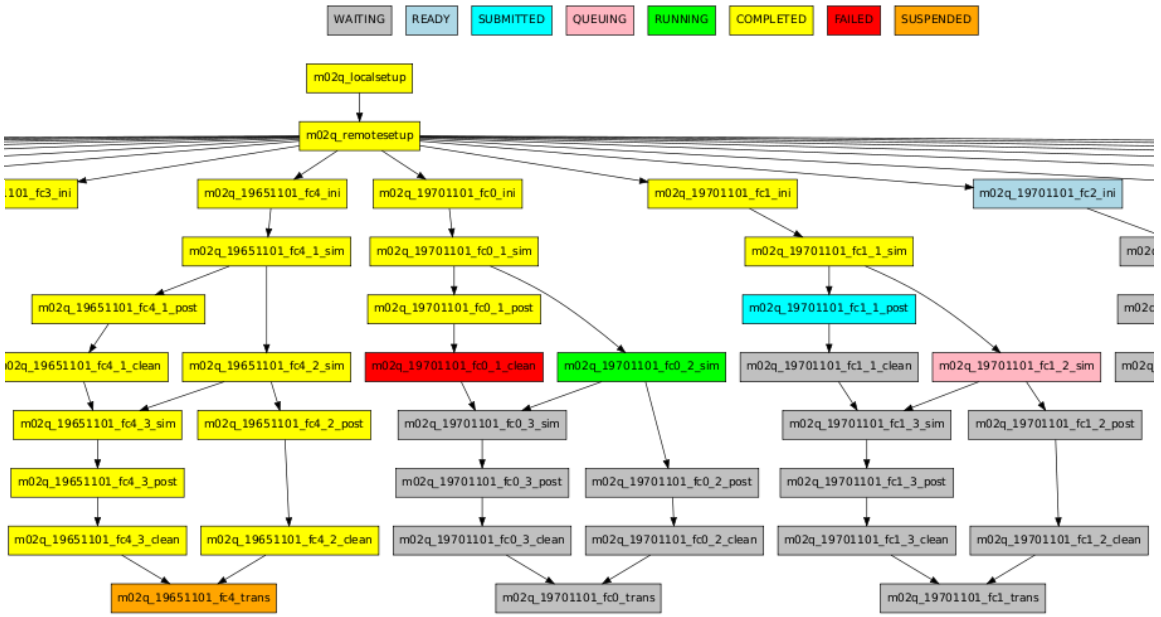


Prodhomme et al. (2015, Clim. Dyn.)

Workflows: Autosubmit

- **Automatisation:** Preparing and running, postprocessing and output transfer, all managed by Autosubmit. No user intervention needed.
- **Provenance:** Assigns unique identifiers to each experiment and stores information about model version, configuration options, etc
- **Failure tolerance:** Automatic retrials and ability to repeat tasks in case of corrupted or missing data.
- **Versatility:** Currently runs EC-Earth. NEMO and NMMB models **on several platforms**.

Workflow of an experiment monitored with Autosubmit (yellow = completed, green = running, red = failed, …)



Case 1: Data streaming for air quality forecasting

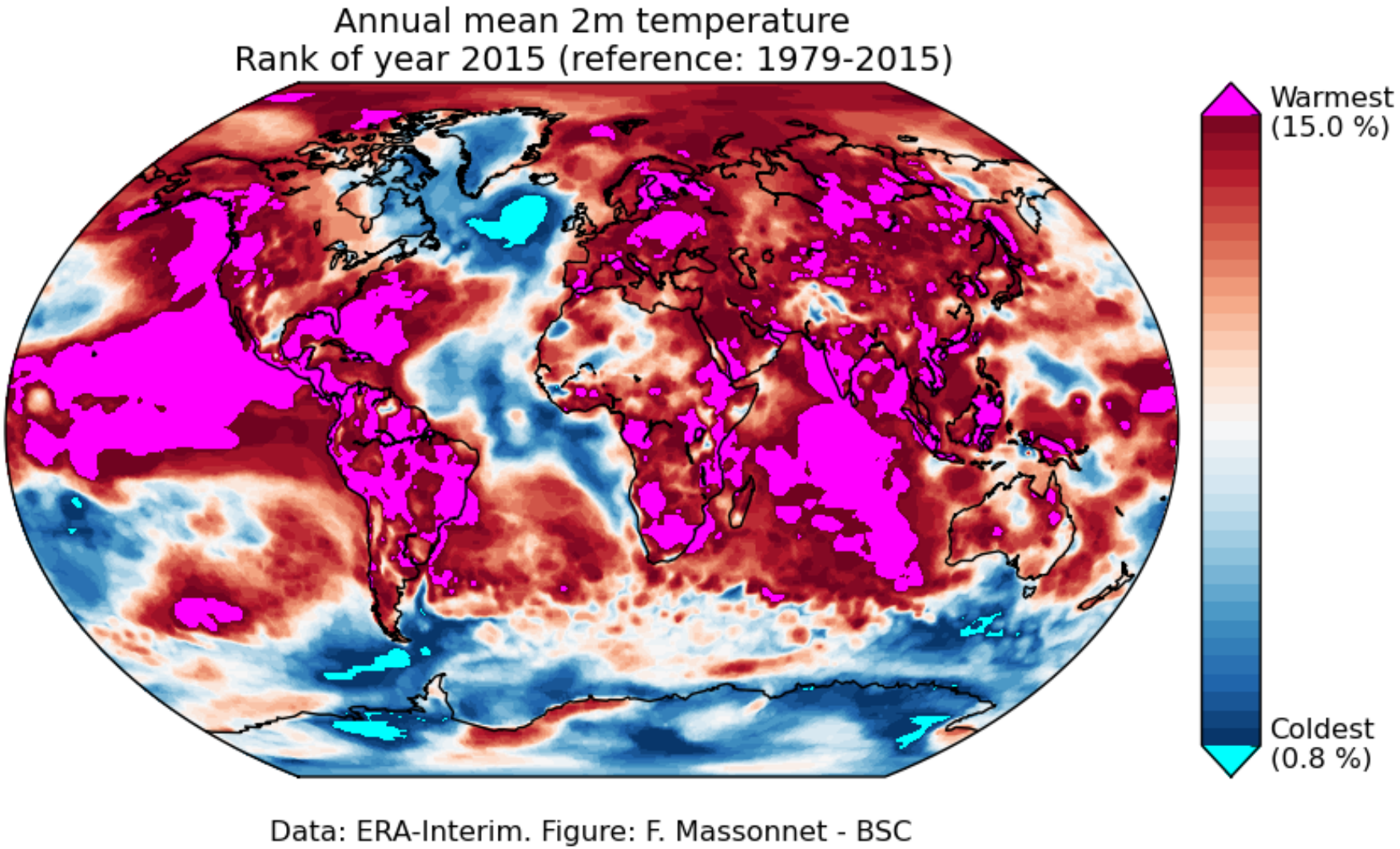
Case 2: Simultaneous analytics and HPC in climate prediction

Case 3: Analytics as a service

Climate change is taking place

Barcelona Supercomputing Center
Centro Nacional de Supercomputación
EXCELENCIA SEVERO OCHOA

Rank of the 2015 annual mean temperature over the last 37 years from ERA Interim.



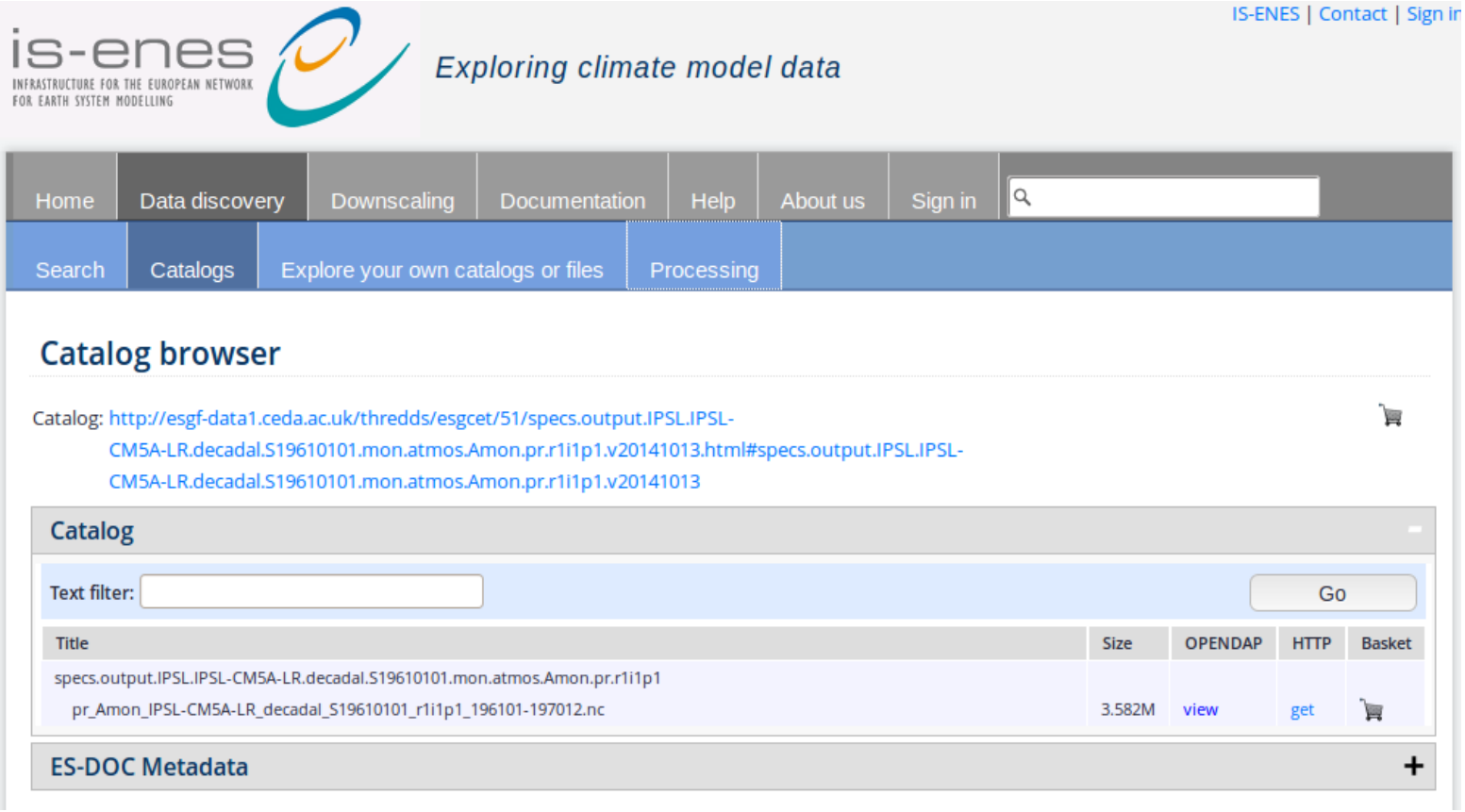
Downstream services

Users need services to respond to their questions.

An example of downstream service http://climate4impact.eu.

This is a key aspect for the success of our research.

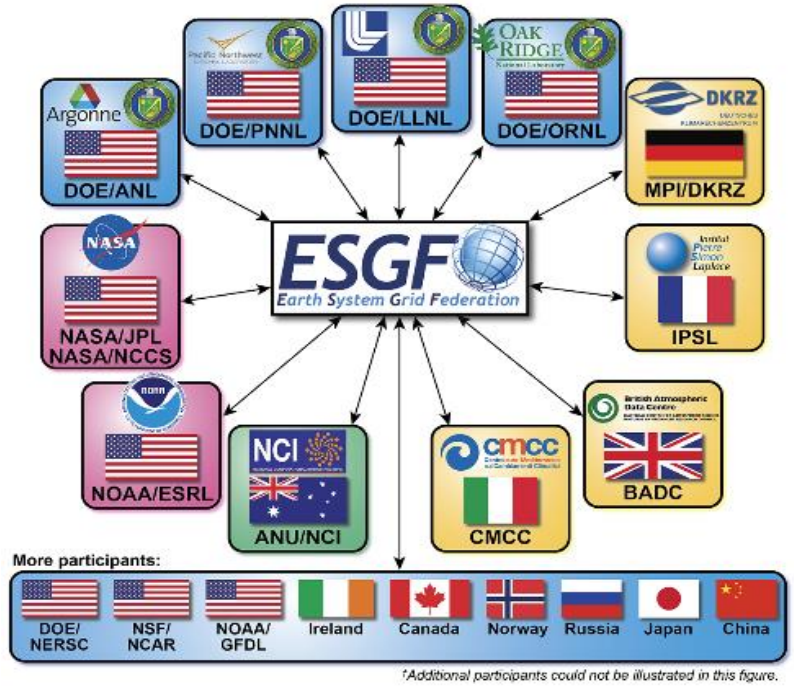
Based on access to **ESGF**.



Data access of downstream services

The Earth System Grid Federation is an open source effort providing a robust, distributed data platform, enabling worldwide access to peta/exascale weather and climate data. ESGF **promotes common formatting, has discovery tools and data indexing**.

EUDAT, a pan-European network, is working on general-purpose additional solutions.



ESGF is organized as worldwide **distributed nodes** and addresses the following aspects

- Having data "stuck" locally and difficult to share among institutions.
- Too complex data indexing and discovery of data repositories.
- Undocumented datasets or lacking metadata.

National and international initiatives

- Spanish Network for weather, climate and air quality data.
  - Initiative started with a workshop in Madrid the 13th of November 2015.
  - Foster the discussion about solutions to issues linked to Big Data in our sectors among the research, operational and industrial Spanish community
  - http://www.bsc.es/projects/earthscience/bigdata
- Research Data Alliance (international initiative with the objective to promote data sharing among all kinds of research communities) **Interest Group in weather, climate and air quality** to address a strong **pressure from users**
  - Workshop held in Barcelona the 11th of February 2016
  - earthsciences-rda-ig@bsc.es, http://rd-alliance.org

Some more open questions

- **Education**: in the era of open data, the community should take advantage of the open education opportunities.
- **Heterogeneity**: how to link to and merge our data with those from communities with larger impact (urban development, arts, social responsibility)?
- **Technology**: how can we make the most of a rapidly evolving technology (heterogeneous nodes, data software, mobile data capture, storage/compression, computing and storage outsourcing)?
- **Awareness**: what are the priority issues for our communities and how can we work better together?
- **Industry engagement:** how can we solve the problem of involving the private sector?